3488–3494 Nucleic Acids Research, 1994, Vol. 22, No. 17

Database on the structure of small ribosomal subunit RNA

Yves Van de Peer, Ilse Van den Broeck, Peter De Rijk and Rupert De Wachter*
Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium

ABSTRACT

The database on small ribosomal subunit RNA structure contains (June 1994) 2824 nucleotide sequences. All these sequences are stored in the form of an alignment based on the adopted secondary structure model, which in turn is corroborated by the observation of compensating substitutions in the alignment. The complete database is made available to the scientific community through anonymous ftp on our server in Antwerp. A special effort was made to improve electronic retrieval and a program is supplied that allows to create different file formats. The database can also be obtained from the EMBL nucleotide sequence library.

CONTENTS OF THE DATABASE

The database on small ribosomal subunit RNA (further abbreviated as SSU rRNA) contained 2824 sequences on June 21, 1994. This number comprises 621 eukaryotic, 77 archaeal, 1997 bacterial, 39 plastid, and 90 mitochondrial sequences. Partial sequences are included only if the combined length of the sequenced segments corresponds to homologous segments in *Escherichia coli* SSU rRNA amounting to at least 70% of the chain length of the latter molecule. All sequences are stored in the form of an alignment and contain the postulated secondary structure pattern in encoded form.

Because of the current size of the SSU rRNA database, it is no longer possible to include a table listing all species for which the SSU rRNA structure is recorded. Instead, only the name of the taxon is given, as well as the number of sequences representing it in the database. Table 1 lists the different eukaryotic taxa. The taxonomic classification of the species is according to Brusca and Brusca (1) for the Animalia, according to Cronquist (2) for the higher plants, according to Ainsworth *et al.* (3) for the zygomycetes and ascomycetes, according to Moore (4) for the basidiomycetes and ustomycetes, and according to Margulis *et al.* (5) for the remaining eukaryotes, viz. the Protoctista.

Table 2 covers the prokaryotic SSU rRNA sequences. The classification is based on the construction of evolutionary trees, explained in more detail in the previous compilation (6). In short, new sequences retrieved from the EMBL (7) and/or GenBank (8) nucleotide sequence libraries are aligned with their presumed closest relative. Evolutionary trees are then constructed by the neighbor-joining method (9), and according to the phylogenetic position observed, the species are assigned to one of the taxa described by Woese and coworkers (10, 11) and our research group (6, 12). In the case of the Bacteria, no hierarchical distinction is made between divisions and subdivisions, such as the α, β, γ, δ, and ϵ subdivisions of the division Proteobacteria, since the subdivisions do not always form together a monophyletic cluster. Moreover, within the Proteobacteria, we also distinguish a γ* division formed by species attributed to the Proteobacteria γ group by Woese (13) but consistently grouped with the Proteobacteria β in our trees (6). For the Archaea, a distinction is made between the divisions Crenarchaeota and Euryarchaeota (14). The latter division is subdivided into 8 subdivisions. Of these, the Methanobacteriales, Methanococcales, Thermococcales, and Methanopyrales correspond to lineages distinguished by Olsen and Woese (14). The remaining subdivisions, the Methanomicrobia, the Halobacteria, *Thermoplasma acidophilum* and *Archaeoglobus fulgidus* are grouped in the subdivision Methanomicrobiales by the latter authors, but do not form together a monophyletic group in our trees (6).

SECONDARY STRUCTURE

Secondary structures encoded in the sequences are based either on the prokaryotic model, which is applicable to Bacteria, Archaea, plastids and mitochondria, or on the eukaryotic model applicable to all Eukarya. Helices are given a different number if separated by a multibranched loop (e.g. helices 9 and 10), by a pseudoknot loop (e.g. helices 1 and 2), or by a single stranded area that does not form a loop (e.g. helices 2 and 32). A single number is given to 50 'universal' helices, which are present in all SSU rRNAs from Archaea, Bacteria, and plastids known to date. They are also present in all known eukaryotic SSU rRNAs except in those of Microspora, where some of these helices are missing. Helices specific to the prokaryotic model are given composite numbers of the form Pa-b, where a is the number of the preceding universal helix and b sequentially numbers all helices inserted between universal helices a and a+1. Helices specific to the eukaryotic model are similarly numbered Ea-b.

Secondary structure models are shown for the prokaryotes, represented by the archaebacterium *Sulfolobus acidocaldarius* (Fig. 1), and the cyanobacterium *Nostoc PCC7120* (Fig. 2) and for the Eukarya, represented by the ciliate *Oxytricha nova* (Fig. 3).

*To whom correspondence should be addressed

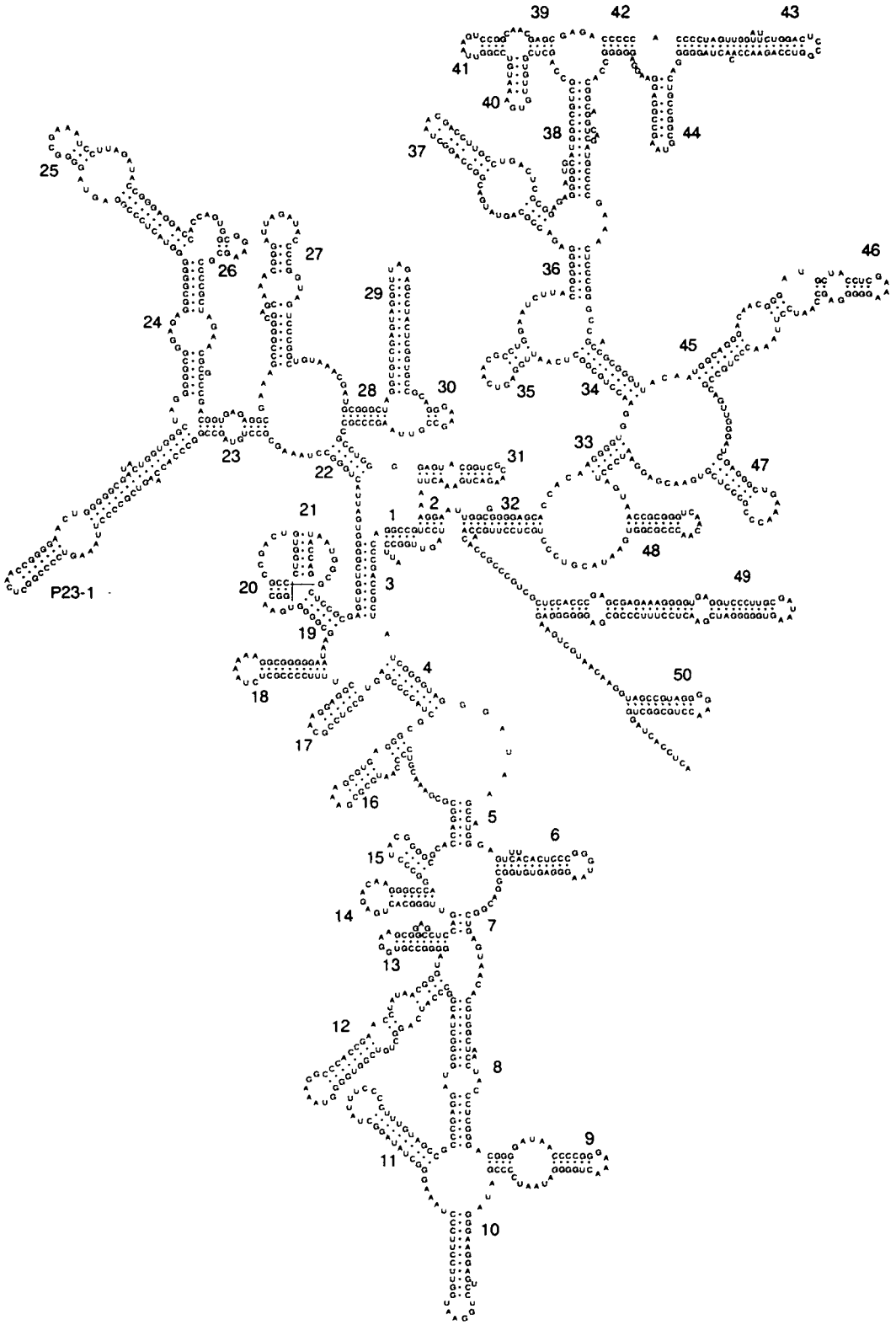

Figure 1. Secondary structure model for SSU rRNA of the archaebacterium *Sulfolobus acidocaldarius*.

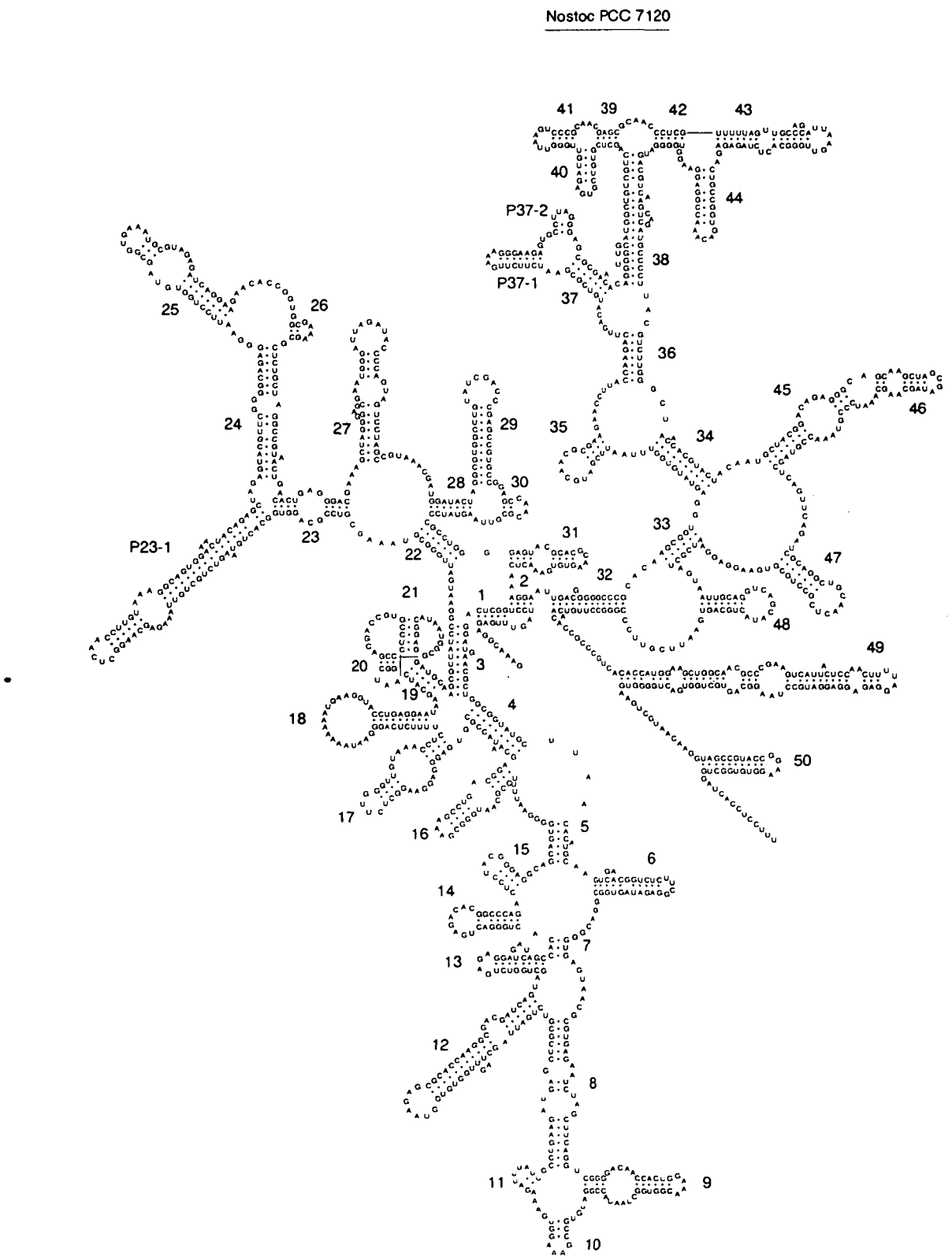

Figure 2. Secondary structure model for SSU rRNA of the cyanobacterium *Nostoc PCC7120*.

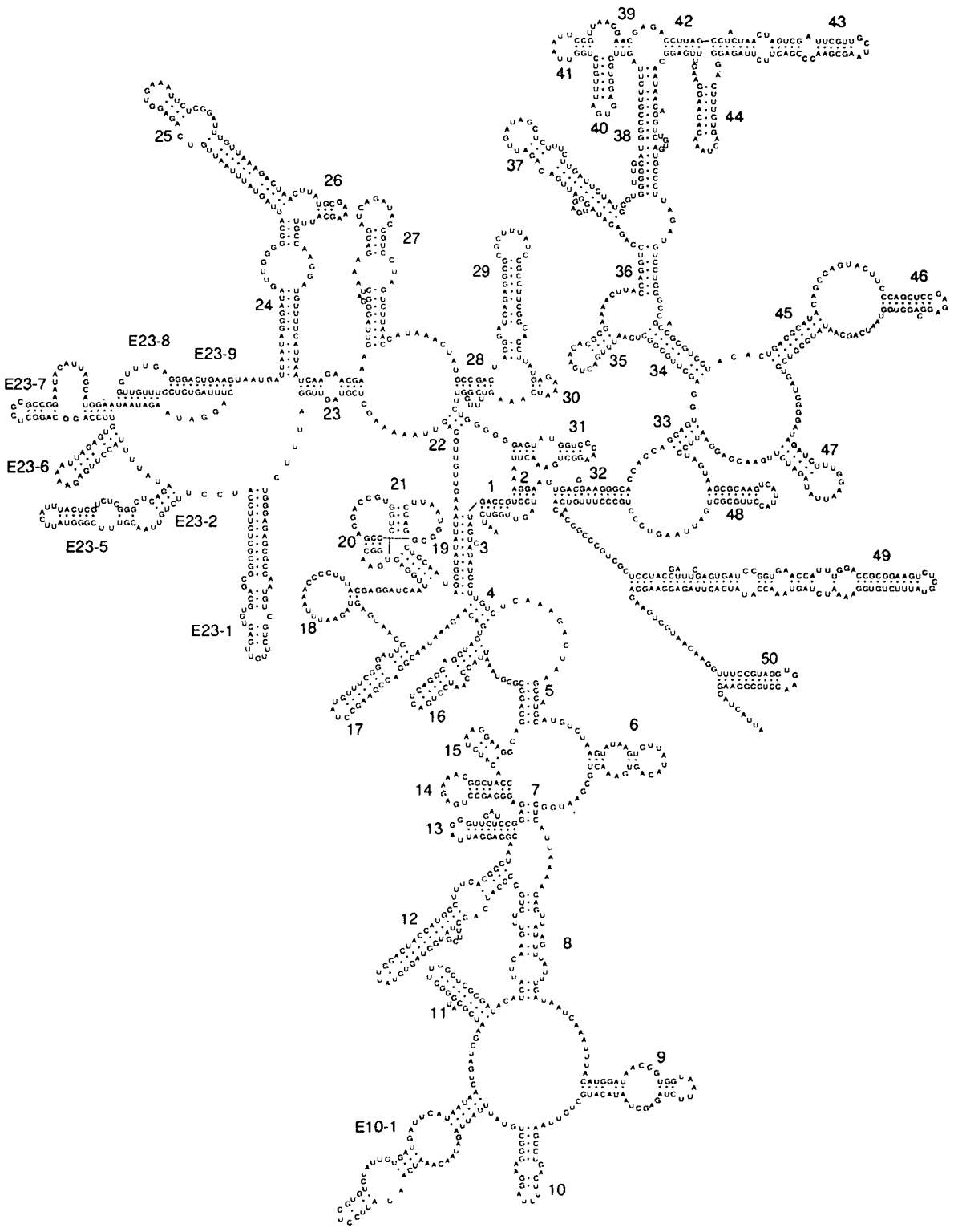

Figure 3. Secondary structure model for SSU rRNA of the ciliate *Oxytricha nova*.

Mitochondrial sequences show extreme variability in length and in the number of helices present. Fig. 4 shows an example of an animal mitochondrial SSU rRNA model, viz. that of the fish *Cyprinus carpio*. Examples of secondary structure models for other mitochondrial SSU rRNAs have been given in a previous compilation (15). Some of these have been subjected to minor changes.

AVAILABILITY OF THE DATA

The SSU rRNA database will be made available through anonymous ftp on the server uiam3.uia.ac.be (143.169.8.1). The files will also be made available to the EMBL nucleotide sequence library for distribution. Researchers who cannot obtain the database through these channels, can request the database or parts

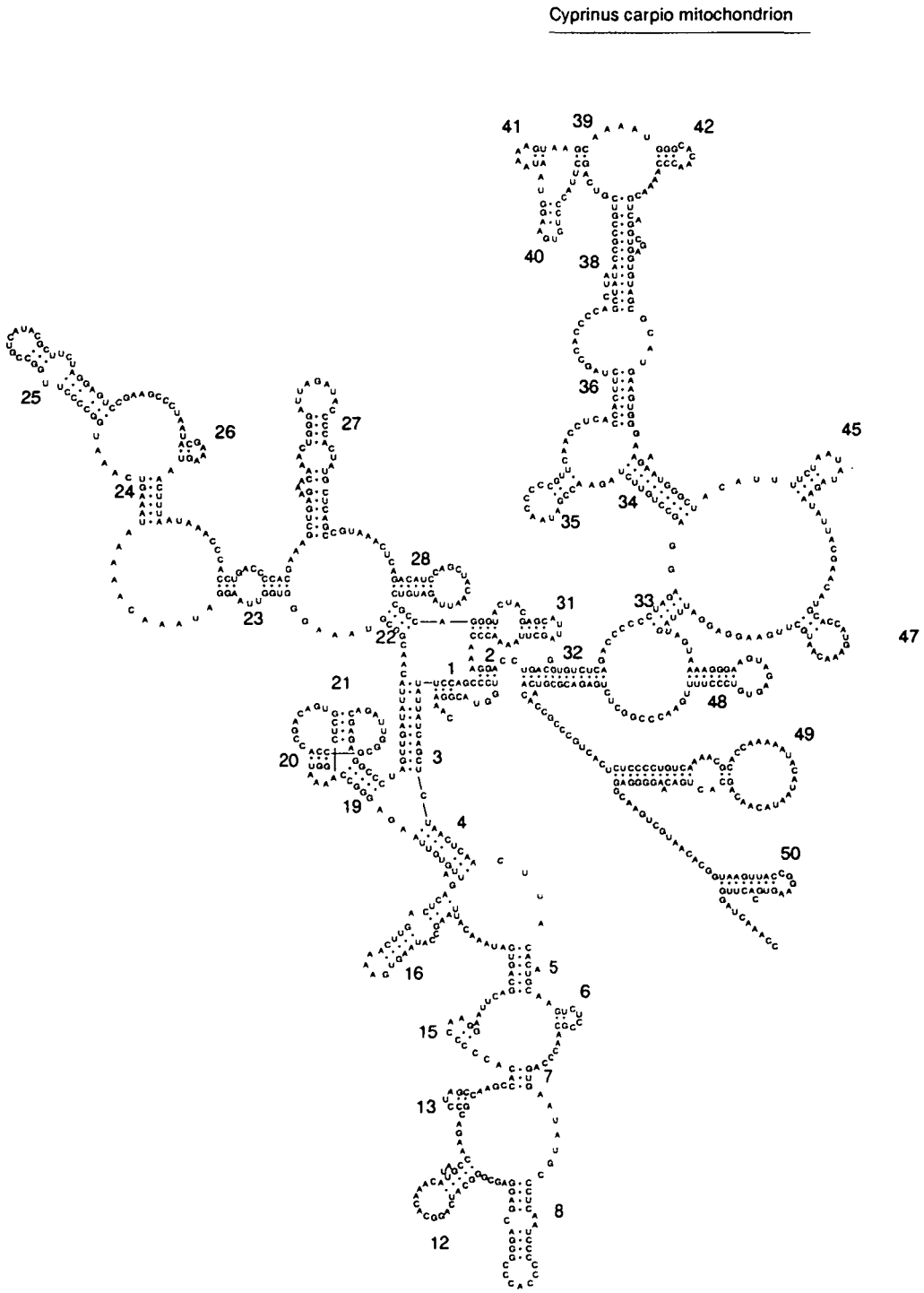

Figure 4. Secondary structure model for mitochondrial SSU rRNA of the fish *Cyprinus carpio* (Osteichthyes).

thereof on magnetic media from the authors. On our server in Antwerp, a file called 'readme' will be present with the contents of database files and directory structure, and a description of available programs for format conversion, alignment editing (16) and phylogenetic tree construction (17). The authors can be contacted by electronic mail to dwachter@reks.uia.ac.be or rrna@reks.uia.ac.be.

In order to simplify access to the data, each SSU rRNA sequence is stored in a separate file. This starts with a few header lines which contain data about the sequence such as the accession number and literature reference, and is followed by the organism name. The sequence comes next. It consists of a range of nucleotide symbols interspersed with gap symbols necessary for alignment. The sequence end is indicated by an asterisk. The beginning and end of secondary structure elements are indicated by insertion of special symbols. The names of these files are produced from the species name by taking characters of the genus and species names. These are preceded by a code describing the phylogenetic group to which the species belongs. This makes it possible to either retrieve specific sequences using the full file name, or to retrieve a set of sequences belonging to a phylogenetic group using wild cards. A program available on the server allows to create different file formats and to integrate several sequences into an alignment.

Users of the database are requested to cite this paper.

ACKNOWLEDGEMENTS

Our research was supported by the BRIDGE programme of the commission of European Communities (contract BIOT-CT91-0294), by the Programme on Interuniversity Poles of Attraction of the Office for Science Policy Programming of the Belgian State (contract 23), and by the Fund for Collective Fundamental Research. Peter De Rijk is research assistant of the National Fund for Scientific Research. We thank Sabine Chapelle for the computer drawings of the secondary structure models.

REFERENCES

1. Brusca,R.C. and Brusca,G.J. (1990) Invertebrates, Sinauer Associates, Inc. Sunderland.
2. Cronquist,A. (1971) Introductary Botany, Harper & Row, New York.
3. Ainsworth,G.C., Sparrow,F.K. and Sussman,A.S. (1973) The Fungi: an Advanced Treatise, Academic Press, New York, Vol. 4A.
4. Moore,R.T. (1988) in Moriarty,Ch. (ed.). Taxonomy putting plants and animals in their place. Royal Irish Academy, Dublin, pp. 61–88.
5. Margulis,L., Corliss,J.O., Melkonian,M. and Chapman,D.J. (eds.) (1990) Handbook of Protoctista., Jones and Bartlett Publishers, Boston.
6. Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res.* **21**, 3025–3049.
7. Rice,C.M., Fuchs,R., Higgins,D.G., Stoehr,P.J. and Cameron,G.N. (1993) *Nucleic Acids Res.* **21**, 2967–2971.
8. Benson,D., Lipman,D.J. and Ostell,J. (1993) *Nucleic Acids Res.* **21**, 2963–2965.
9. Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
10. Woese, C.R. (1987) *Microbiol. Rev.* **51**, 221–271.
11. Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol.* **176**, 1–6.
12. Van de Peer,Y., Neefs,J.-M., De Rijk,P., De Vos,P. and De Wachter,R. (1994) *System. Appl. Microbiol.* **17**, 32–38.
13. Woese,C.R. (1991) In Selander,R.K., Clark,A.G., Whittam,T.S. (eds.). Evolution at the Molecular Level. Sinauer Associates, Inc., Sunderland, pp. 1–24.
14. Olsen,G.J. and Woese,C.R. (1993) *FASEB J.* **7**, 113–123.
15. Neefs,J.-M., Van de Peer,Y., De Rijk,P., Goris,A. and De Wachter,R. (1991) *Nucleic Acids Res.* **19**, 1987–2015.
16. De Rijk,P. and De Wachter,R. (1993) *Comput. Applic. Biosci.* **9**, 735–740.
17. Van de Peer,Y. and De Wachter,R. (1993) *Comput. Applic. Biosci.* **9**, 177–182.

Table 1. List of eukaryotic taxa represented in the database and number of their representatives

Kingdom Animalia[a] Phylum	Class	No. sequences[b] N	M
Placozoa		2	
Porifera	Calcarea	2	
	Demospongiae	2	
Cnidaria	Anthozoa	2	
	Cubozoa	1	
Ctenophora		2	
Platyhelminthes	Trematoda	12	
	Turbellaria	3	
	Uncertain affiliation	1	
Nematoda	Secernentea	6	2
Acanthocephala	Archiacanthocephala	1	
Arthropoda	Chelicerata	3	
	Insecta	5	3
	Malacostraca	17	
	Maxillopoda	1	
Pentastomida	Pentastomata	1	
Mollusca	Bivalvia	11	
	Gastropoda	1	
	Polyplacophora	1	
Echinodermata	Echinoidea	2	
Chaetognatha		1	
Chordata	Agnatha	4	
	Amphibia	17	2
	Aves	2	2
	Chondrichthyes	3	
	Mammalia	7	41
	Osteichthyes	3	6
	Reptilia	4	
	Cephalochordata (Subph.)	1	
	Urochordata (Subph.)	2	
Total:		118	58

Kingdom Fungi Subphylum	Class	No.sequences[b] N	M
Zygomycotina	Zygomycetes	14	
Ascomycotina	Discomycetes	6	
	Hemiascomycetes	50	7
	Loculoscomycetes	16	
	Plectomcycetes	23	2
	Pyrenomycetes	10	1
	Uncertain affiliation	1	
Basidiomycotina	Heterobasidiomycetes	21	
	Hymenomycetes	7	
Ustomycotina	Ustomycetes	11	
Total:		159	10

Kingdom Plantae Division	Class	No. sequences[b] N	M	P
Bryophyta	Bryopsida	1		
	Marchiantiopsida		1	1
Magnoliophyta	Magnoliopsida	35	3	11
	Liliopsida	2	5	2
Pinophyta	Cycadopsida	1		
	Ginkgoopsida	1		
	Gnetopsida	1		
	Pinopsida	6		
Total:		47	9	14

Kingdom Protoctista[c] Phylum	Class	No. sequences[b] N	M	P
Apicomplexa	Coccidia	22		
	Hematozoa	35	2	

Table 1. (*cont.*)

Kingdom Protoctista[c] Phylum	Class	No. sequences[b] N	M	P
	Uncertain affiliation		4	
Bacillariophyta	Bacillariophyceae	5		
	Coscinopiscophyceae		4	
Chlorarachnida		4		
Chlorophyta	Charophyceae	9		
	Chlorophyceae	54	2	10
	Prasinophyceae		4	
	Ulvophyceae		1	
Chrysophyta	Chrysophyceae	4		2
	Uncertain affiliation	2		
Chytridiomycota		7		
Oomycota		3		
Ciliophora		29	5	
Conjugaphyta	Conjugatophyceae	4		
Cryptophyta		5		2
Dictyostelida		1		
Dinoflagellata		10		
Euglenida		1		6
Eustigmatophyta	Eustigmatophyceae	1		
Granuloreticulosa		2		
Microspora		10		
Phaeophyta		4		2
Plasmodial slime molds	Myxomycota	1		
Prymnesiophyta		9		1
Rhizopoda	Lobosea	7		
Rhodophyta		19		2
Xanthophyta		1		
Zoomastigina	Amebomastigota	3		
	Choanomastigotes	2		
	Diplomonadida	6		
	Kinetoplastida	23	4	
	Parabasalia	1		
Total:		297	13	25

[a]The Metazoan taxa are listed in the same order as they appear in (1).
[b]The number of sequences listed in the database is larger than the number of species, because for certain species multiple SSU rRNA sequences have been determined, usually by different authors. The sequences are not necessarily identical because they may have been determined for different varieties or strains of a species, or for different genes of the same organism. The number is listed for sequences of nuclear (N), mitochondrial (M), and plastid (P) origin.
[c]The Protoctist phyla and classes are ordered alphabetically.

Table 2. List of prokaryotic taxa represented in the database and number of their representatives

Bacteria Division	No. sequences[a]
Chlamydiae	3
Cyanobacteria	38
Fibrobacter	15
Flavobacteria and relatives	133
Fusobacterium and relatives	27
Gram positives and relatives, low G+C	588
Gram positives and relatives, high G+C	235
Green non sulfur	4
Green sulfur	4
Planctomyces and relatives	7
Proteobacteria Alpha	254
Proteobacteria Beta	89
Proteobacteria Gamma	237
Proteobacteria Gamma*	97
Proteobacteria Delta	51
Proteobacteria Epsilon	76
Proteobacteria uncertain	8
Radioresistant micrococci and relatives	27
Spirochetes	73
Thermotogales	4
Uncertain affiliation	27
Total:	1997

Archaea Division	Subdivision	No. sequences[a]
Euryarchaeota	Archaeoglobales	1
	Halobacteria	11
	Methanobacteriales	16
	Methanococcales	5
	Methanomicrobia	24
	Methanopyrales	1
	Thermococcales	2
	Thermoplasma	1
Crenarchaeota		16
Total:		77

[a]The number of sequences listed in the database is larger than the number of species, because for certain species multiple SSU rRNA sequences have been determined, usually by different authors. The sequences are not necessarily identical because they may have been determined for different strains, or for different genes of the same strain.